2016 Employment Projections Technical Paper

Published August 2016

Washington State Employment Security Department
Dale Peinecke, Commissioner

Labor Market and Performance Analysis
Cynthia Forland, Ph.D., Director

Prepared by

Alex Roubinchtein, Projections and Statistical Analyses Manager
Bruce Nimmo, Economic Analyst

This workforce product was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration.

Two stage projections process

In stage one, industry projections are created. Six steps are required to complete this stage. In stage two, industry projections are converted into occupational projections. Five steps are required to complete this stage. In this report we will describe each stage in the order it occurs.

Industry Projections

Data used

- Covered employment time series
- Global Insight forecasts

Software used

Since last year, the primary software used for forecasting has been R-software (R). R is an open source object oriented language with advanced statistical and optimization features. It allows programmers to operate directly on vectors and matrices. This creates significant advantages over sequel-based languages, like SAS, when producing occupational projections.

Step 1. State level aggregated industry forecast

Data preparation

Initial covered employment at the county level was aggregated into 42 industry groups (cells), presented in Table 1:allcodes.xlsx. Forty cells were aggregated for nonfarm employment, one for agriculture and one for private households. The cells for nonfarm employment are closely associated with employment related cells from the Global Insight model. However, to meet state employment projections requirements and Occupational Employment Statistic (OES) definitional requirements, some cells were disaggregated for state projections. For example, we disaggregated transportation equipment to aerospace and other transportation equipment. The state and local government cell was disaggregated into three cells: government education, hospitals and other government. Two industries related to the information sector were also disaggregated.

We transformed some codes from the Global Insight model in order to match them with codes used in state projections. Due to these transformations, 40 state cells obtained matching relationships with Global Insight national forecasts. Two state cells, agriculture and private households, do not have related national forecasts.

A crosswalk between 4-digit NAICS codes, Industry Control Total (ICT), aggregated series codes and common combined codes can be found at: allcodes.xlsx. As can be seen in the *allcodes.xlsx* file, aggregated series do not in all cases represent an aggregation of ICT codes. The main reason is that aggregated series reflect commonly used definitions from the Current Employment Statistic (CES) classification system, while ICT codes reflect industry definitions used in the OES system. To match CES and OES systems, we created detailed combined codes. The combined codes are used to match aggregated series forecasts with detailed ICT forecasts.

The Global Insight model uses data with quarterly frequencies. In contrast, state forecasts use data with monthly frequencies. To make national forecasts usable as regressors for state forecasts they must be interpolated from quarterly into monthly frequencies. To achieve this we used the **denton-cholette method** from the R-library **tempdisagg**.

The main procedure

The main industry projections procedure consists of two parts: 1) non-indexed, for all series; and 2) indexed by series. Three additional libraries were used for industry forecasts: **forecast**, **dynlm** (dynamic linear model), and **foreach**. The non-indexed procedure involves the import of all data and the defining of all necessary subsets, objects and seasonal dummy and time variables for different time intervals. Objects are held for later use in each series when indexing and cross indexing occur.

For each of the 40 state cells, which have regressors (Global Insight interpolated forecasts), we use the following four types of models:

- Exponential smoothing: innovations state space autoregressive model with optimized selection of smoothing parameters (criteria minimum Mean Absolute Percent Error (MAPE)).
- Auto ARIMA: optimized selection of parameters of ARIMA, seasonal ARIMA, period of seasonality, etc. with regressors (criteria: AIC (Akaike's information criterion)) - this is probably the most sophisticated single equation model available.
- Naive regression model with only seasonal dummies and time (linear trend) as regressors.
- Dynamic linear regression model which includes regressors (the same as for Auto ARIMA), seasonal dummies and linear trend.

The exponential smoothing and naive models are autoregressive and only use historical employment time series to forecast employment. The auto ARIMA and dynamic linear regression models can include independent variables (regressors).

The state space method offers a unified approach to a wide range of models and techniques. In general, it includes equations for unobserved states and includes observation equations. Unobserved states (such as level, growth and seasonality) can be subject to change with time. Since the model can account for such changes, it is called **innovative**. The general model can be described as follows:

Let $x_t = (l_t, b_t, s_t, s_{t-1}, ..., s_{t-m+1})'$, be a state vector, where l_t - stands for level; b_t - for growth; and s_t - for seasonality. State space equations can be written in the form:

$$y_t = w(x_{t-1}) + r(x_{t-1})\epsilon_t$$
$$x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t$$

where ϵ_t - error terms with mean zero and variance δ^2. The equation $\mu_t = w(x_{t-1})$ is a one step ahead forecast for the states y_t - observed numbers (employment in our case). Other parameters are defined by the type of mode. For instance, models with multiplicative errors use $r(x_{t-1}) = 1$ resulting in $y_t = \mu_t(1 + \epsilon_t)$. Thus, relative errors for multiplicative models are represented by $\epsilon_t = (y_t - \mu_t)/\mu_t$. As can be seen in the state space model, the term "dynamic" refers to states, rather than to observed numbers as in traditional descriptions.

In the **forecast** package, similar state space models for 30 exponential smoothing variations are subject to internal optimization. In our model specifications, for this year's forecast, we chose to allow a damping parameter as a variable. In previous years this was not the case.

You can find more technical details about the models which were used in the R-package **forecast** at: http://robjhyndman.com/papers/automatic-forecasting/.

The next two types of models are traditional regressions with dynamic, not one step ahead, forecasts. The dynamic linear regression model is presented in the form:

$$y_t = c + a * g_t + d * t + s_1 + ... + s_{11} + \epsilon_t$$

where observed employment numbers, y_t, are the linear function of intercept c, endogenous global insight forecasts, gi, and 11 seasonal dummies, s. If the intercept was not used there would be 12 seasonal dummies. The naive regression model is the same with the exception of the component related to regressors, $a * gi_t$.

For each time series and each model, two forecasts are produced:

- one based on a full sample; and
- one based on a 24 month hold-out sample.

For the full sample forecast, we used all available historical data from January 1990 to June 2015 for parameter estimations. We then forecast for the period from July 2015 to December 2024. Estimations for the hold-out forecast are based on historical data from January 1990 to June 2013 and then forecast from July 2013 to June 2015. As a result of this method, for each time series we have eight forecasts.

We use the following optimization procedure to define weights for combining four full forecasts and we calculated eight mean absolute percentage errors. For estimations on full samples for each model class:

$$MAPE = (\sum_{t=1}^{306} |\epsilon_t|/y_t)/306,$$

where 306 is the number of months between January 1990 and June 2015.

For the hold-out sample forecasts, the MAPE's were calculated as:

$$MAPE = (\sum_{t=283}^{306} |y_t - for(y_t)|/y_t)/24$$

where $for\,(y_t)$ represents a forecast on a hold-out sample (training-set in R terminology). We defined the optimum four weights $z = (z_1, z_2, z_3, z_4)$ for combining forecasts for four model classes by solving the problem, find $z = (z_1, ..., z_4)$, for which:

$$\sum_{i=1}^{4}(\sum_{t=1}^{306} |\epsilon_t|/y_t)/306 * z_i + \sum_{i=1}^{4}(\sum_{t=283}^{306} |y_t - for(y_t)|/y_t)/24 * z_i \rightarrow min$$

The combined optimum forecast is defined as $\sum_{i=1}^{4} for(y_t^i) * z_i$.

For two series without regressors we use the same procedure, but only with three types of models. The naive model and dynamic linear regressions become equivalent and the last is excluded from the process. Also, regressors are excluded from the auto ARIMA model.

Outcomes of the main procedure

The main procedure produces five forecasts for each time series: four models plus a combined optimum forecast. We repeat this procedure for log transformed series and thus potentially have 10 forecasts for each series.[1] There are two options for combining the forecasts with historical data:

- use the actual historical data; or
- use fitted historical data.

We use both options. The result being that we can have up to 20 combinations of forecasts with historical data. These combinations are used as regressors for aggregated workforce development area (WDA) forecasts and state detailed forecasts.

[1]Estimations for some models can fail for a variety of reasons. The chance for failure increases for unstable series with small numbers involving some zeros. To avoid interruptions in loop processing, for failed series, we use **tryCatch** loops, rather than the default **do** loop. An error handling function prints I.D.'s for all failed series. Also, using the **foreach** loop , rather than the more common **for** loop allows us to have all of the successful forecasts in output lists as well as identification of all failed series.

Formal adjustments to industry forecasts

Adjustments are applied to all combinations of forecasts and historical data. An adjustment is a useful procedure for smoothing results. We used the concept of **stability controls for dynamic systems** as our smoothing method. The variance of historical employment growth over 12 months[2] was used to define confidence intervals for projected employment variances. We also arbitrarily established the lower and upper confidence limits at 0.96 and 1.04. The intervals represented the lower number between the historical confidence and the established limit. For each time point, if projected numbers fell within established intervals, it stayed. Otherwise, limits were applied. This process was used as the main mechanism for adjusting models.

Formally the adjustment procedure for each of the series y_t, $t = 1, 2, ..., 420$[3] can be described as follows.

Twelve month growth rates calculated as:

$$gr_i = y_t/y_{t-12}, \quad i = 1, ..., 408, \quad t = 13, ..., 420$$

A total of 294 growth rates represent historical data, while another 114 represent forecasted data. We calculate 95% confidence intervals for historical growth rates (*high* and *low*) and average growth rates (*mean*). In this current version of adjustments, we are using only *high* and *mean.* To make the adjustment formulas more understandable we introduce two new variables: *adj = high - mean* and *base = max(1,mean).* Then adjustments to the forecasted growth rates gr_i, $i = 295, ..., 408$ are produced by subsequent application of upper and lower limits as follows:

$$gradj_i = gr_i \; if \; gr_i < min(1.04, (base + adj)) \; otherwise \; gradj_i = min(1.04, (base + adj))$$

then

$$gradj_i = gradj_i \; if \; gradj_i > max(0.96, (base - adj)) \quad otherwise$$
$$gradj_i = max(0.96, (base - adj))$$

where, 0.96 and 1.04 are arbitrarily selected numbers and can be subject to change.

The order of applying upper and lower adjustments is irrelevant since values will be unaffected.

Adjusted forecasts are simply produced by multiplying the last year of available historical data by adjusted growth rates. Then the adjusted forecasts are combined with historical data. Adjustments are applied to each available series, up to 20, resulting in up to 40 forecast options.

Selection of aggregated state forecasts

At this stage, we select just one from up to 40 state aggregated series combined forecasts. The selected series will be used as regressors in later steps. It is possible that a selected series represents a linear combination of a few forecasts. However in this round of projections we stayed with just single series selections. The selection is an **informal process** and is based on calculated average annual growth rates for periods used for the current round of projections. For this round of projections, they were: 2015Q2 - 2017Q2; 2014 - 2019 and 2019 - 2024. The growth rates are calculated from aggregated monthly series to proper frequencies (quarterly or annual). The following considerations are used in the informal selection of forecasts:

- historical growth rates for the entire period and the last 10 years;
- the latest aggregated long-term employment forecast from the Office of Financial Management (OFM) and short-term forecast from the Economic and Revenue Forecast Council (ERFC);
- previously published forecasts: our forecasts, OFM and ERFC forecasts;

[2] Twelve month (or over-the-year) growth rates are used to avoid the impact of stable seasonality.
[3] Combined series include 420 months (from January 1990 to December 2024).

- smoothness of transitions between the last month of historical data and the first month of the forecasted data;
- general knowledge of underlying trends in specific industries;
- avoiding extreme growth and decline rates.

We try to select forecasts with growth rates close to those used by OFM and ERFC. We do this unless we have convincing evidence to the contrary or OFM and ERFC forecasts are inconsistent between themselves or have significant differences with previously produced results. ERFC forecasts are used for budgetary planning purposes and require the use of adaptive controls. Consequently, forecasts should be updated often to reflect the best and most current data. In such cases, up-to-date data takes priority over long, stable forecasting time periods. Our forecasts are mainly used for career development. Prospective students need forecasts that are stable for medium-range time periods. Frequent updates of forecasts in such cases would be disruptive. In other words, for prospective students, frequently updated forecasts lose practical value.

An example of applying general knowledge of underlying industry trends to specific industries can be seen in our prerecession analyses of productivity trends. Specifically, in the construction industry, our analysis demonstrated that a combination of high employment growth coexisted with a high rate of declining productivity. The declining productivity was compensated for by large price index growth. The combination of high employment growth, low productivity and high prices could not last indefinitely and pointed to a high probability of a downward correction. Therefore, our employment projections for construction growth were more pessimistic. In fact this type of correction happened during the great recession and created a large drop in construction employment. This drop was the largest among all major industry sectors.

A similar situation occurred in the forecasting of aerospace employment. The delay of Boeing's Dreamliner aircraft combined with high demand created an artificial boom in aerospace employment trends. Our projections were more in line with normal aerospace long-term declining cyclical trends. In both cases our declining trends were subject to strong criticisms. Subsequent events affirmed the practice of applying knowledge of underlying trends. The artificial aerospace conditions eventually ceased and declining cyclical trends continued.

An assortment of special cases occurred in this projections cycle. One series, private households, received special handling after it experienced a large code change. The code change broke the series. Out of 41 selected forecasts, 34 were optimum forecasts. In 21 optimum forecast selections, the selections were combined with historical data and in 13 cases were combined with fitted historical data. Of the 41 selected forecasts, the exponential smoothing model was selected in 3 cases and two ARIMA and two naive model forecasts were selected. The above numbers justify the complicated process of using optimum combinations on the four types of model.

Among the 41 selected series, 16 resulted from log transformed unadjusted forecasts, 6 from log transformed adjusted forecasts, 14 were original (i.e., not log transformed) unadjusted forecasts and 5 were original adjusted forecasts. As a result, all 4 types of transformations are present in the selected series.

Selected series forecasts are used in the following three independent, but related steps.

Step 2. Draft of state level aggregated benchmarked forecast

The actual historical covered employment numbers for the last 18 months are combined with noncovered employment provided by the Current Employment Statistics (CES) unit. These numbers are rolled up to two base point forecasts: average annual for the year 2014 and average quarterly for the second quarter of year 2015 (2015Q2). This procedure is called **benchmarking**. Unlike benchmarking by the CES unit, we do not use any wedging or other adjustments to incorporate code changes. Thus, our benchmark numbers can be slightly different from CES numbers. The growth rates from selected forecasts are applied to benchmarked base numbers. The result is that we produce three required points for industry forecasts: 2017Q2, 2019 and 2024. Since shares of non-covered employment can be significantly different for different cells, forecasted employment shares for benchmarked employment can be different from shares for covered employment. One of the unpleasant results of this difference is the need for balancing the hierarchy by areas (between state and totals of workforce development areas) and industries (between aggregated and detailed industries). Balancing of covered employment forecasts does not automatically lead to balanced benchmarked forecasts.

The results of benchmarked forecasts are rolled up to create multi-level tables that are somewhat comparable with CES tables. The table is submitted to regional economists, state agencies specified in state law and the Economic Revenue Forecast Council. Feedback from these groups is requested.

Step 3. Detailed state level forecasts

Main procedure

We used selected aggregated forecasts as regressors for state detailed forecasts. Nine detailed series are the same as aggregated and one series, education, represents totals of two aggregated series. For these series, aggregated forecasts (or their combination) were used as detailed forecasts. For all other 265 series, we use the same **Main procedure** as in **Step 1** for series with regressor. Unlike **step 1**, we did not produce forecasts for log transformed series. As a result we had 20 forecast options for each series. Potentially it could be less if the fitting of some models for some series failed, but this did not occur at the state level.

Adjustment of state detailed forecasts

Detailed forecasts should be adjusted to match aggregated forecasts. As was discussed in the data preparation part of step 1, matches between detailed and aggregated series are made at the level of combined codes. For cases where detailed codes equal aggregated or their combination, results are the same and thus already matched. However for the remaining 29 combined series, matches should be made between 265 detailed forecasts and 32 aggregated. In prior years, this matching was for the most part done manually. This year for the major adjustments we used a process which produced weighted detailed forecasts for each detailed series which most closely matched projected growth rates for aggregated series at the level of the combined series.

For the adjustment process we used three of five available classes of models: exponential smoothing, auto ARIMA and optimum combination. The main reason for using three classes was to limit the sizes of the respective optimization tasks. While dynamic regression and naive models are not directly included, their contributions were reflected in optimum combinations.

Let's use indexes s, m and v for identifying short-term (2015Q2 -2017Q2), medium-term (2014-2019) and long-term (2024) employment projections. The 32 aggregated forecasts were further rolled up to 29 combined series. The average annual forecasted growth rates g_s , g_m and g_l were calculated for each of the combined series k. Let's define subset S_k of detailed series for each combined series k . Let's use index i for detailed series ($i \in S_k$), index j for models options in each detailed series ($j = 1, ..., 12$) and index t ($t = 1, ..., 420$) as a time index for all series with monthly frequencies. The optimization problem for each of the combined series can be written as follows.

Find the weights $w_{i,j}$ of aggregation detailed forecast options for each detail series from the following conditions:

$$w_{i,j} \geq 0, \quad i \in S_k, \quad j = 1, ..., 12$$

$$\sum_{j=1}^{12} w_{i,j} = 1, \quad i \in S_k$$

$$\sum_{i \in S_k} \sum_{j=1}^{12} w_{i,j} (|(\sum_{t=328}^{330} y_{i,j}^t / \sum_{t=304}^{306} y_{i,j}^t)^{0.5} - 1 - g_s| +$$

$$+ |(\sum_{t=349}^{360} y_{i,j}^t / \sum_{t=289}^{300} y_{i,j}^t)^{0.2} - 1 - g_m| + |(\sum_{t=409}^{420} y_{i,j}^t / \sum_{t=349}^{360} y_{i,j}^t)^{0.2} - 1 - g_l|) \rightarrow min$$

where $y_{i,j} = (y_{i,j}^1, ..., y_{i,j}^{420})$ vectors of employment numbers for detailed forecasts i and option j.

In simpler words, we define the normalized not negative 12-th dimensional vectors of weights for each detailed series, which will weight the forecast options in order to minimize average absolute differences for three periods in annual average growth rates between aggregated combined series weighted detailed forecasts and given aggregated series forecasts.

After weights are determined, the weighted detailed state forecasts for each detailed series i ($i = 1, ..., 265$) are calculated as:

$$y_i^{f,t} = \sum_{j=1}^{12} w_{i,j} * y_{i,j}^t$$

The procedure immediately above was implemented in MS Excel using Add-Ins Solver. One of the reasons for using Excel rather than R was immediate visibility of results. In this case, Excel made it easier to make small interventions. The interventions were made when weighted detailed forecasts had significant changes between the last month of historical data and the first month of forecast data, or for unreasonable growth rates. The interventions were limited to modifications of weights. Such modifications can correct such problems, but leave the objective function values close to optimum.

Step 4. Local workforce development area (WDA) forecasts

Main procedure

The procedure for producing and formally adjusting local level aggregated and detailed forecasts are the same in a mathematical sense. The only difference is that for detailed series we did not use the log transformation option. We used state aggregated and detailed forecasts from the previous steps as regressors for WDA's aggregated and detailed forecasts. The same **Main procedure** as in **Step 1** for series with regressor was used in this step. There are three possible outcomes from this procedure for each series:

4.1 - All options did not fail and thus we have 40 outputs for aggregated series and 20 for detailed series.

4.2 - All options failed and thus we do not have any forecasts.

4.3 - Some options failed and we have numbers of outputs less than 40 or 20.

If all options failed,[4] we assign statewide growth rates for such series. For outcomes 4.1 and 4.3, we use formal adjustment procedures close to those described in **Step 1**.

Adjustments of WDA's forecasts

The purpose of this step is to select for each available forecast, among all available options, in each area, a weighted forecast. As in previous steps, we want a forecast that produces growth rates for periods of interest most close to ones at the state (regressors) level. Adjustments on this step are completely formal, conducted in R and exclude interventions.

To keep our indexes somewhat consistent with the model, described in the previous step, let's define g_s, g_m and g_l as short-term (2015Q2 -2017Q2), medium-term (2014-2019) and long-term (2024) average annual growth rates employment projections for each state forecast (aggregated or detailed). Let's also keep using index t ($t = 1, ..., 420$) as a time index for all series with monthly frequencies and j as index for forecast options for each available series i. For outcome 4.1, it will be 40 for aggregated series and 20 for detailed. The numbers will be less if some forecast failed (outcome 4.3). Let's define the numbers of not failed forecasts for each series i as n_i. Then the optimization problem for each of the available series i can be written as follows.

[4] Some local (aggregated and detailed) series might not exist (i.e., have zero covered employment). They also can be interpreted as failed series.

Find the weights w_j of aggregation for forecast options from the following conditions:

$$0 \leq w_j \leq 1, \quad j = 1, ..., n_i$$

$$\sum_{j=1}^{n_i} w_j(((\sum_{t=328}^{330} y_j^t / \sum_{t=304}^{306} y_j^t)^{0.5} - 1 - g_s)^2 +$$

$$+ ((\sum_{t=349}^{360} y_j^t / \sum_{t=289}^{300} y_j^t)^{0.2} - 1 - g_m)^2 + ((\sum_{t=409}^{420} y_j^t / \sum_{t=349}^{360} y_j^t)^{0.2} - 1 - g_l)^2 \rightarrow min$$

where $y_j = (y_j^1, ..., y^4 20_j)$ vectors of employment numbers for option j.

After weights are determined, the weighted forecasts for each series i are simply calculated as:

$$y_t^f = \sum_{j=1}^{n_i} w_j * y_j^t$$

Step 5. Draft of aggregated industry forecasts

Formally adjusted aggregated industry forecasts are benchmarked in the same manner as described in **Step 2**. After benchmarking, numeric discrepancies between the state and WDAs are resolved through an informal process. Since discrepancies are normally very small, due to formal adjustments, for the most part state numbers are just slightly modified to meet WDA totals. In some cases the inverse is required and WDA numbers are subject to adjustments to meet state totals.

Minimal informal interventions are possible at this stage of the projections process. Interventions are based on known discrepancies between state and local area trends. In this round of projections, an intervention was applied to electronic shopping and mail-order houses. The extreme state growth rates in this industry are mainly due to historical trends in King County and thus needed to be smoothed. The smoothing needed to occur since state level numbers are used as regressors for areas.

After adjustments are made, industry tables for each WDA are created the same way as in **Step 2** and submitted for internal review by regional economists.

Step 6. Final adjustments and output of industry employment projections

This step is mainly informal and involves processing feedback from state agencies and regional economists. There are two types of responses. Each response type is processed differently:

1. Responses based on expected, event-based information; and
2. Responses on the level of suggestions, related to major trends.

Event-based information is related to expected closures and layoffs or expected new hirings due to business expansions or relocations. A very conservative approach is followed and only events with high degrees of certainty are used. For example, an adjustment was made to the alumina/aluminum industry based on copies of warn notices from the Economic and Revenue Forecast Council. Adjustments were made for the state and affected area(s). Some response-based adjustments were also made to South West Washington industry forecasts. The event-based adjustments are coordinated between aggregated and detailed forecasts. In addition, we incorporate informal interventions for aggregated series from **step 5** to detailed series.

Suggestions related to major trends are evaluated based on available data and underlying economic trends. For all suggestions we develop responses. We either provided reasons for rejecting suggestions or inform suggestion sources that suggestions will be incorporated into projections. Suggestions are mainly incorporated in the most conservative manner. The main way to incorporate trend-based suggestions at the state level is by returning back to **Selection of aggregated state forecasts** in **step 1** and repeat all the subsequent steps for the affected industries. It is also possible to modify models for affected industries. For suggestions related to local areas we return back to **step 4**.

After all adjustments are made and all aggregated and detailed projections are benchmarked, informal adjustments are required to meet the following balancing-requirement. These balancing-requirements must be met for each of the three projected time periods (2017Q2, 2019 and 2024):

- For each industry, totals for local areas for aggregated and detailed industries should be equal to state numbers; and
- For each area, balances between detailed and aggregated forecasts should be achieve at the aggregated combined series level.

Satisfying the above two conditions leads to balances between state and local areas forecasts at the combined series aggregation level.

While **step 6** processes may seem complicated, for the most part they are not difficult or time consuming after all automated adjustments have been made. Discrepancies are normally not large and mainly can be handled by either a bottom-up approach where state totals are made equal to areas totals or by using a top-down approach with proportional adjustments to local areas numbers to meet state totals. For a few series with multiple cross-match-adjustments at the combined series level, the process is more complicated.

The industry projections process produces two major outputs:

1. Aggregated industry projections for the state and all areas, which are rounded to the closest 100 and rolled up to create a multi-level table that is somewhat comparable with CES tables (as in **Step 2**).

2. Industry control total (ICT) files for the state and all areas, which are not rounded.

The aggregated projections output is published and is used for analyses in projections reports, but is not used for producing occupational projections. The non-rounded ICT output is used in subsequent steps for producing occupational forecasts. ICT industry projections numbers, rounded to integers, are also published.

Occupational Projections

Data used

Occupational employment projections result from the conversion of industry employment into occupations. These conversions are based on occupation/industry ratios (i.e., staffing patterns) from the Occupational Employment Statistics (OES) survey. The survey is conducted by the Labor Market and Performance Analysis (LMPA) division of the Employment Security Department (ESD) in cooperation with the U.S. Bureau of Labor Statistics (BLS). The full OES survey is conducted over a three-year cycle. One-third of the survey is completed each year. Occupational estimations and projections are subject to the limitations of the OES survey. The survey includes nonfarm employment and agriculture services, but excludes noncovered employment, self-employment and unpaid family members, major agriculture employment (except services), and private households.

The sample for the OES survey is designed for metropolitan statistical areas (MSA). From the perspective of statistical accuracy for occupational projections, this level of aggregation is most appropriate. However, for different applications like the Training Benefits Program, we used WDA aggregation levels for regional details. The direct use of OES staffing patterns for WDAs can create significant bias for a variety of reasons.

In this round of projections, the data source for the creation of staffing patterns was almost entirely raw survey data. The majority of data comes from BLS final (i.e., not preliminary) files. These files include employment, employment distributions by wage intervals, final weights and indicators showing if original survey responses or imputed responses are used. Response imputations come from other similar in-state areas or from other states. Imputations can have a significant influence on the OES-based staffing patterns.

The process of selecting staffing patterns for each industry and area includes calculating industry totals from raw files. Totals are calculated for weighted employment with imputation and without imputation. Totals are then compared with Industry Control Totals (ICT) for the base year periods 2014 and 2015Q2. Our preference is to use data without imputations, but in some cases they do not represent significant shares of employment in ICT files. In such cases, either samples with imputations or substituted staffing patterns are used. These substitutions are mainly introduced using statewide staffing patterns. In some cases substitutions come from other similar in-state areas. Staffing patterns can create significant bias for industries with high shares of noncovered employment, which were not part of the survey (e.g., religious organizations).

For a few industries, combined staffing patterns were used between areas. This mainly occurred for the King County and Snohomish County WDAs. This was a necessary step because King and Snohomish counties were combined in the OES survey sample. National staffing patterns are used as a last resort and only one was used for private households.

Some problems, however, were unavoidable and significantly influenced final occupational estimations and projections. For example, doctors can be employed by clinics or hospitals, but often are employees of independent associations or are self-employed. For this reason, staffing patterns for medical institutions were bound to be biased. Also noteworthy this year was the limited use of some of the results from the 2012 OES green supplement survey for agricultural industries. The green supplement survey allowed us to create staffing patterns for agriculture, based on weighted sample responses. We also used older survey responses to account for a major missing reporter which has been missing from the latest surveys.

To manage this process, we added two new columns to our ICT files. The new columns indicate the **type** and **area of origination** for staffing patterns. For instance, if an original staffing pattern is used, the area of origination will be the same as the area for industry employment. If an original is not used, the area of origination might be the numeric indicator zero (0) for statewide substitution or the numeric indicator 45 for the combination of King and Snohomish counties, etc.

Occupational projections use the following national inputs:

- Self-employment and unpaid family worker ratios.
- Replacement rates for each occupation.
- Change factors, which we modify.

Step1. Making staffing patterns with selected change factors

The new **type** and **area of origination** columns are used to create base staffing patterns for all areas. **Type** is used to define the source file for selection, while a combination of **area of origination** and **ICT code** is the key for selecting records from source files. The source files are:

- Raw survey data with and without imputations.
- Extracts from the 2012 OES green supplemental survey for agricultural industries.
- National staffing patterns (used only for private households).
- Older raw sample responses for a large missing employer.

Combined indexes in the **areas of origination** column are split apart for the selection process of original area codes. Thus, area 45 will be split into a 4 and a 5. For both areas, all available ICT codes are repeated. After selection, combined codes are restored by a simple summarization of numbers for each available industry and occupation.

Each selected vector $a_{i,j}^{v}$, where $i = 1, .., m$, index for occupations, $j = 1, ..., n$, index for industries and ($v \in V$) index for areas (can be the number or combination of numbers) is normalized:

$$\sum_{i=1}^{n} a_{i,j}^{v} = 1, \ for\ all\ industries\ j\ and\ areas\ v.$$

The combination of vectors $a^v_{i,j}$ is often called a **matrix of staffing pattern**. It represents the normalized structures for the distribution of industry employment between occupations for the base period(s). The vectors $a^v_{i,j}$ are matched with extended ICT files by the **area of origination** and industries. After this match, the **area of origination** column is not used for further calculations and thus can be dropped from further formal descriptions. Instead we now can use the index of actual area v for WDA's from matched files.

Let's now define the index for base and projected periods as $t, t = 1, .., 5$ and for this round of projections it represents the years 2014, 2015Q2, 2017Q2, 2019 and 2024. The base staffing patterns are used for the years 2014 and 2015Q2 (t=1,2). For other periods, patterns are modified with the incorporation of limited change factors.

Change factors $c_{i,j}$ come from national data. They predict expected changes in occupational shares for each industry over ten years. The reliability of change factors tends to be low because unlike industry employment, there are no historical time series for occupational employment. Due to the lack of historical trends upon which to base future changes, BLS uses researchers' expectations about structural occupational changes within industries to create change factors. Within this BLS process, there is a high degree of subjective judgment. This is especially true since change factors must be developed for each occupation within an industry. Occupational outputs are very sensitive to these change factors. It is very important to evaluate the adequacy of change factors before use. Incorrect change factors can drastically increase errors in projections. We created change factors for a limited numbers of cells. This was done where state historical series were available and were consistent with the suggested change factors from national files. In such cases, we used the most conservative estimations. Change factors reflect expected changes over ten years and staffing patterns for projected periods must be modified accordingly:

$$c^2_{i,j} = (c_{i,j})^{0.2}, \quad c^5_{i,j} = (c_{i,j})^{0.5}, \quad c^{10}_{i,j} = c_{i,j}.$$

For the two base periods, change factors are not used. The value for all missing change factors can be assumed to be one and modified staffing patterns are calculated as:

$$a^{v,t}_{i,j} = (c_{i,j})^t * a^v_{i,j} \quad t = 1, .., 5,$$

where $a^v_{i,j}$ are the staffing patterns for the base period. Staffing patterns for each period, industry and area need to be normalized to totals of 1.

Step 2. Calculation of occupational projections

The results from the previous calculations, for each component $x^t_{j,v}$ of the original ICT vectors, in each time period, give as output normalized vectors for occupational distributions $a^{v,t}_{i,j}$. The base occupational employment for each period is simply calculated as:

$$e^v_{i,t} = \sum_{j=1}^{n} a^{v,t}_{i,j} * x^t_{j,v}, \quad i = 1, .., n, \quad t = 1, .., 5 \quad v = 0, ..., 12$$

due to:

$$\sum_{i=1}^{m} a^{v,t}_{i,j} = 1 \quad for\, each\, j,\, v\, and\, t\, we\, have \sum_{i=1}^{m} e^v_{i,t} = \sum_{j=1}^{n} x^t_{j,v}.$$

The totals of occupational employment for each area in each point of time equals the totals of industry employment.

The numbers for the base period 2015Q2 represent distributions of industry employment between occupations according to normalized staffing patterns. Often, they are called staffing patterns. These staffing patterns are convenient for publications, but need to be normalized for any calculations outside base periods or with modified ICT files.

Step3. Calculations of self-employment and unpaid family members

Raw self-employment ratios s_i for each occupation come from national data. Based on these ratios, we calculate unadjusted estimated self-employment totals for each area for the base year period 2014 as:

$$se_l = \sum_{i=1}^{m} s_i * e^v_{i,1}$$

We use estimated numbers of self-employed from the American Community Survey to adjust national self-employment ratios for each area. Let's define the survey numbers for each areas as $self_l$. The ratio of adjustment is defined as $ratio_l = self_l/se_l$. The ratio is assumed to be the same for all periods and in this way, adjusted numbers of self-employed for each area v and occupation i are defined as:

$$ase^v_{i,t} = se^v_{i,t} * ratio_l$$

Then the total of occupational employment is defined as:

$$et^v_{i,t} = e^v_{i,t} + ase^v_{i,t}$$

Step4. Adding openings due to replacement or separations

Replacement and separation rates comes from national data. Traditionally we used only one type of replacement rate, net replacements. This year we used an experimental alternative separation rate. From a mathematical point of view, calculations are the same for the two rates. Let's define the rates as r_i. Then the openings due to replacement or separations for each occupation for each period defined as follow:

$$rep_{i,v} = \frac{et^v_{i,b} + et^v_{i,f}}{2} * r_i,$$

where $et^v_{i,b}$ and $et^v_{i,f}$ are employment totals for the beginning and end of the period. We calculate replacements for periods between 2015Q2 and 2017Q2, 2014 and 2019, 2019 and 2024.

Step 5. Making final outputs

The final outputs include the following results. Calculations are rounded to integers and aggregations to totals for 2 and 3 digit SOC levels:

- Total occupational employment estimations $et^v_{i,t}$ for all five periods.
- Average annual growth rates for three periods: 2015Q2 and 2017Q2 , 2014 and 2019, 2019 and 2024.
- Average annual number of openings due to growth $gr_{i,v}$ for each period, which are calculated by subtracting starting points from end points and then dividing the results by the number of years in the period (2 or 5).
- Average annual openings due to replacements $arep_{i,v}$ calculated by dividing $rep_{i,v}$ by the number of years in the period.
- Total openings due to growth and replacements are calculated as follows:

$$tot_{i,v} = max((arep_{i,v} + gr_{i,v}), 0)$$

We initially rounded employment estimations and then aggregated them to total 2 and 3 digit SOC codes. In this way, results are additive for each column. However,the above formula for calculating total openings introduces non-additivity into the calculations. As a result, the aggregated level of total openings might not equal the total of growth plus replacement.

Some detailed occupational employment estimations are suppressed due to confidentiality. The suppression is introduced after aggregations and normally is not reflected in the aggregated results.